

STUDIES IN RECENTLY DEVELOPED GROUP-FORMING PROCEDURES IN TAXONOMY AND ECOLOGY

A. V. HALL

(Bolus Herbarium and Department of Botany, University of Cape Town)

ABSTRACT

Some important properties of methods of forming groups in taxonomy and ecology are briefly discussed. An illustration is given of the inappropriateness of using Heterogeneity Analysis, a new group-forming procedure, and perhaps also Information Analysis, for grouping samples rather than unique items. A means of detecting classes and super-classes in dendrograms is described. Using these concepts, several tests of the use of heterogeneity functions for taxonomic and ecological grouping problems are described and evaluated. An outline is given of the computing procedures used in this work. It is concluded that Heterogeneity Analysis shows considerable promise for group-forming studies.

THEORETICAL CONSIDERATIONS

Recently, new systems of forming polythetic sets in Taxonomy and Ecology, using group heterogeneity, were briefly outlined (Hall, 1967). In the matrices of trial linkages of each member or group with every other, fusions are chosen which give the least heterogeneous groupings. After each fusion, the heterogeneity values between the new group and others are calculated in a way that uses all the raw data for each of the individual members taking part in the trial links. The majority of other group-forming procedures involve averaging methods when more than two members are compared, leading to significant losses of information.

Information analysis (Lance and Williams, 1966a) is closely related to Heterogeneity Analysis in that all the raw data about the members of a trial grouping are considered in calculating the value for choosing the best fusion. However, only two-state data coded as 0 or 1 may be used with the Information Statistic. Attributes which are zero for all members of a subset are ignored. This would give rise to difficulties for Taxonomy, where similarities based on either of the two states may be equally important, and where subdivision of the attribute cannot be allowed because of consequent extra weighting. The Information Statistic is clearly more suited to the presence/absence data of Ecology. Common

absences of a given species in a subset of plots are not significant and may be quite appropriately ignored. Even in Ecology, however, the Information Statistic would not seem to be ideal in its present form, as abundances are not taken into account. Often important ecological changes are expressed chiefly in terms of altered frequencies, especially in species-poor areas.

1. *Difference between grouping Samples and Unique Items:* Both Information Analysis and Heterogeneity Analysis appear to be unsuited to grouping samples, and should be reserved for sets of unique items such as the species in a genus or the vegetation types in a given region. This was evident in an application of Heterogeneity analysis to a vegetation study in the present work, and may be illustrated by the following example.

At an advanced stage in a grouping study, three subsets A, B and C remain to be linked. Each represents a different kind of vegetation. It happens, however, that there are thirty sample plots for each of the vegetation types A and B, and only one for C. (See Fig. 1). Of all three, C is the most distinctive kind of vegetation. If A and C were competing for linking on to B, the single sample plot of C might cause a smaller increase of heterogeneity (or lowering of Information content) than the thirty samples of A. In this way, C, although representing a more peculiar vegetation type, would take precedence over A in linking to B.

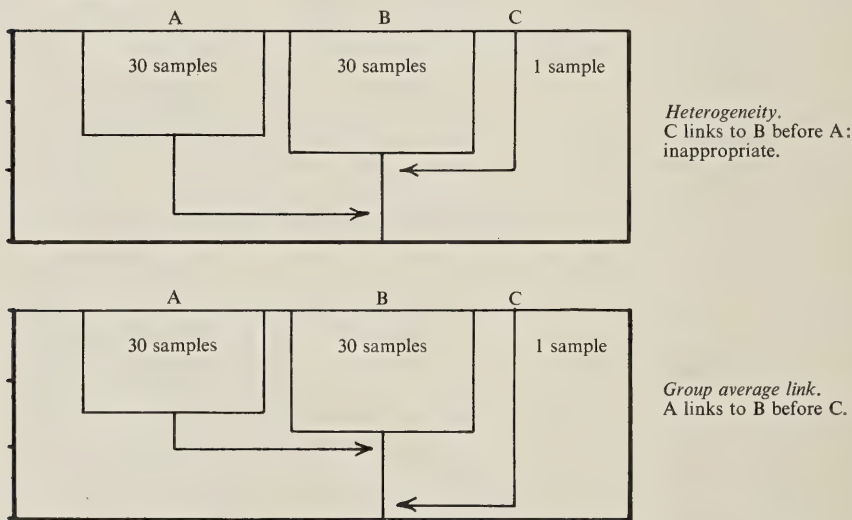


FIG. 1. Dendrograms illustrating the linkage of a poorly represented but distinctive class (C) by the heterogeneity and group average systems.

This fallacious result would be avoided by programming the test for linkage as: how different are the classes of such things as A, B and C? In such a method, the value to be obtained in each comparison is the average of the results of linking each sample in a group to all, respectively, in the other group. This is known as Group Average Linkage (Lance and Williams, 1966b; Sokal and Michener, 1958). The results given by such a study may then be interpreted to give classes, each consisting of a group of one or more samples (see below). These classes may then be fitted together to form the most homogeneous groups possible, using Heterogeneity or if desired Information Analysis, with data that is averaged no more than within each class.

2. *Detecting Classes and Super-Classes:* Broadly speaking, classes may be recognised when the proportional extensions of diversity for the groups in a given link, and for subsequently added subsets, are abruptly larger than usual. This denotes a change from the compact grouping conditions one expects within a class, to the much more diverse situation when classes are brought together. This is illustrated in Fig. 2, where for subset E, $c^1/c > b^1/b$, and a similar condition exists for subsets F, G and H.

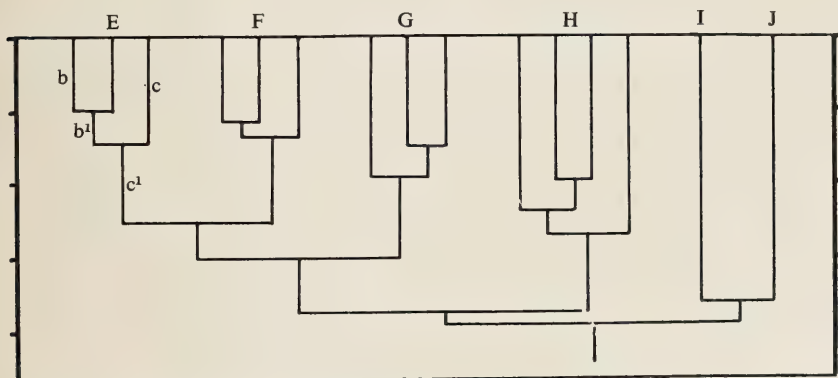


FIG. 2. Dendrogram for illustrating interpretation procedure. (For explanation, see text.)

A special case is shown in Fig. 2 by I and J, which individually are "highly compact", in fact single-membered, "groups". Together they form such a relatively poor group that they are best considered as distinct classes in their own right. Super-classes may be detected by re-applying these principles to the dendrogram in the same way, except for regarding the already-formed classes as single members.

Using personal judgement, this system gave generally reasonable results in the present studies. It is hoped to test it further in a numerical form suitable for the computer. Heterogeneity values, which seem to give a direct measure of compactness, are particularly suited to this treatment. It may be noted that the Phenon Line system of Sokal and Sneath (1963, p. 251) would not seem to be quite adequate for the case in Fig. 2: a level distinguishing E, F and G would cut across subset H in an unsatisfactory way.

TESTS OF HETEROGENEITY FUNCTIONS

1. *Test of Heterogeneity Function for Two-state Data:* The function for two-state data having a form that is insensitive to whether the group is odd- or even-numbered (Hall, 1967), was used in this study. For the j th of the p attributes, the actual number of rare states a_{jr} is divided by the value a_{jrh} that would be obtained for the rare states of an imaginary, maximally heterogeneous case, having the same number of included members. The function may be written as follows:

$$H_t = \frac{1}{p} \sum_{j=1}^p \frac{a_{jr}}{a_{jrh}}$$

The two-state data compiled for the 40 South African taxa of *Eulophia* (Orchidaceae) for a previous study (Hall, 1965), were used in testing this function (see Fig. 3). 93 attributes were used in the description of the taxa.

In spite of this large number of attributes, there were several cases where at a given heterogeneity level, more than one pair of groups could be chosen for linkage. Of a total of 24 possible alternatives, many appeared to involve only minor changes in the positions of members, leaving the subsequent structure unaffected. Three pathways that resulted in different positions of larger groups were investigated. The case that gave the highest total homogeneity (on average, the most compact grouping), was chosen for detailed study.

The total homogeneity values were in fact quite similar: 27.79979, 27.64826 and 27.54613. It remains to be investigated whether a better strategy would be to weight each level by a factor based on the number of included members: it would seem that the homogeneity (compactness) of the larger groups may be more important than for the smaller in seeking the best structure. In the present study, the sums of the values for the last five links show much the same relationships between the three alternative cases as the totals for all the linkages. On neutral grounds also, the groupings in the numerically best case appear to be more logical.

The structure shown in Fig. 3 is similar in many respects to that given in the

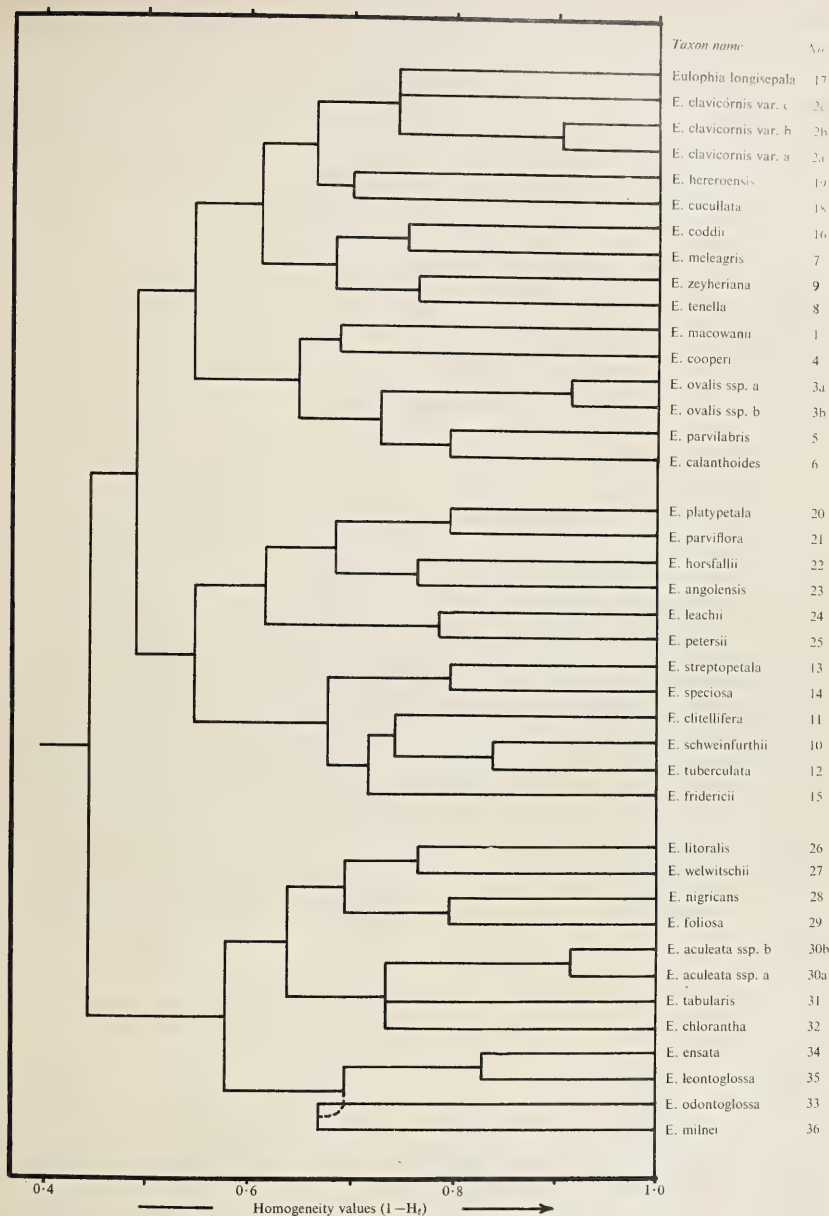


FIG. 3. Homogeneity dendrogram of the South African taxa in the genus *Eulophia*, based on 93 two-state coded attributes. The taxon numbers are the same as those used in a previous study (Hall, 1965).

former study with this data (Hall, 1965, p. 54), using the simple matching coefficient S_{sm} of Sokal and Michener (1958), with the weighted pair-group linking system (Sokal and Sneath, 1963). According to the scheme proposed earlier in this paper, the S_{sm} dendrogram shows four quite clear groups (taxa 1-6, 7-15, 16-25, 26-36), each of which is not easily subdivided. In the heterogeneity (H_t) dendrogram three main groups may be detected (in Fig. 3, taxa 17-6, 20-15, 26-36), two of which may be subdivided (17-8, 1-6; 20-25, 13-15).

Comparing the two dendrograms, both have a generally similar structure in the groups that have less than three or four members. Among the larger groups the *Eulophia litoralis* cluster (26-36) is about equally distinctive, and in both cases has the same contents linked in a somewhat similar way. The arrangement in other large groups differs somewhat and in the majority of cases the heterogeneity dendrogram seems more appropriate.

The rather distinctive *E. streptopetala* group (13-15), formerly appearing as scarcely distinguished from *E. meleagris*, *E. tenella* and *E. zeyheriana*, is suitably isolated in the H_t structure as a sub-group linked with the somewhat similar *E. platypetala* group (20-25). *E. meleagris*, *E. tenella* and *E. zeyheriana* are more appropriately placed elsewhere (cluster 17-8). Similarly, the set containing *E. ovalis* and appropriately, *E. macowanii*, is rather better placed in the H_t analysis as a distinct sub-group. *E. hereroensis* and *E. cucullata*, however, do not seem as well positioned as in the S_{sm} dendrogram. In both dendrograms, *E. clavicornis* var. *nutans* (var. c) links in a way showing Specific rather than Varietal rank. This may be a result of the inefficiency of two-state coding of the attributes. Finally the close grouping of *E. longisepala* with *E. clavicornis* in the H_t dendrogram is an improvement on their position in the former study.

2. *Comparative tests of the Heterogeneity Functions for Taxonomy using different types of Data:* Using the form of the Heterogeneity function for two-state data that was given in the previous test, the formula for both quantitative and two-state data taken together becomes as follows (c.f. Hall, 1967):

$$H_{tq} = \frac{1}{p} \sum_{i_t=1}^{p_t} \frac{a_{jtr}}{a_{jtrh}} + \frac{1}{p} \sum_{j_q=1}^{p_q} \frac{s_{jqn}}{S_{qn h}}$$

Here, the subscripts t and q refer to two-state and quantitative data respectively; s_{jqn} refers to the standard deviation of an attribute j for a set of n members and $S_{qn h}$ to the value for an imaginary group with the same number of members, together having maximal heterogeneity. For a data maximum of 100, a 3-membered maximally heterogeneous group would have attribute values thus:

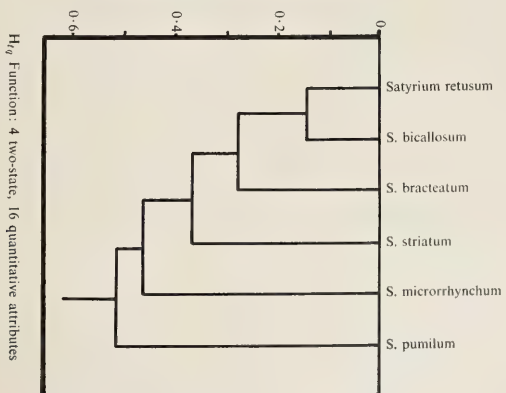
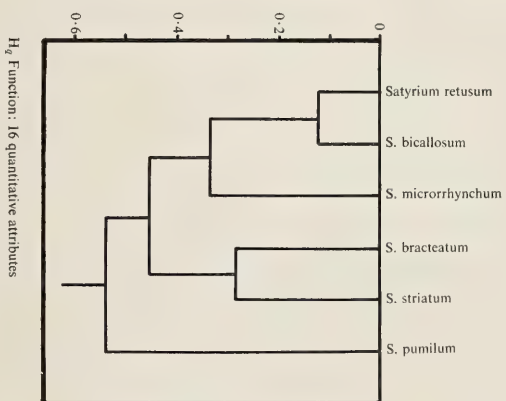
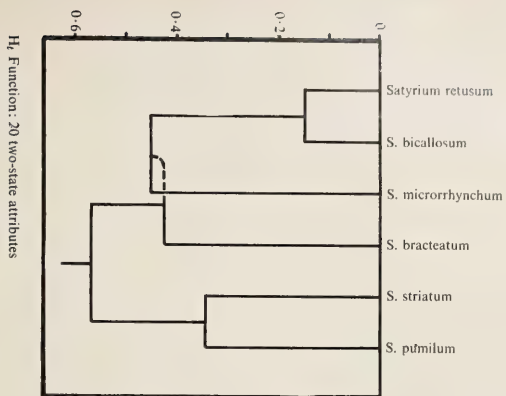


FIG. 4. Dendrograms showing the grouping of six species of *Satyrium* using three different forms of Heterogeneity Analysis.

0 100 100; or thus: 0 0 100. The function for quantitative heterogeneity alone may be given as

$$H_q = \frac{1}{p} \sum_{j=1}^p \frac{s_{jn}}{S_{nh}}$$

Heterogeneity is converted to homogeneity by subtracting the values of H_t , H_q , or H_{tq} from 1.

For the tests, values for 16 quantitative (measurements and scores) and four two-state (presence/absence) attributes were obtained for six species of *Satyrium* (Orchidaceae). For the test of the quantitative function, the four two-state attributes were omitted. For the two-state function, the 16 quantitative attributes were re-written so that values in one half of the range were coded in one state, those in the other half in the alternative state. The results are shown in Fig. 4. *S. microrrhynchum* and *S. retusum* may be interchanged in the H_t dendrogram, giving an alternative pathway but identical subsequent structure.

Comparing the dendrograms, the two-state structure, in either of its equally possible forms, is the most different, and departs significantly from expected groupings based on the grounds of personal judgement. Particularly unsatisfactory is the relatively high level of the link of *Satyrium pumilum*, very probably the most peculiar species present, which on occasion has been regarded as belonging to a distinct genus, *Aviceps*. There is probably too little information given by the two-state coding. A better structure is given by the H_q function with 16 quantitative attributes. Here the position of the lanky, large-leaved and very small-flowered *S. microrrhynchum* seems questionable. The groupings are almost fully satisfactory in the H_{tq} dendrogram with four two-state and sixteen quantitative attributes. A sub-group formed by *S. bracteatum* and *S. striatum* might just be preferable to the linking one after the other shown in the dendrogram. Perhaps this may occur when other attributes are used in a proposed further study, which will include all taxa in *Satyrium*.

When such a study is made, *S. microrrhynchum* may well appear in another group, such as with *S. parviflorum*, leaving *S. pumilum* at an appropriately isolated level. This possibility illustrates the importance of including all related taxa when investigating category boundaries and optimal arrangements.

3. *Test of Homogeneity Analysis for Vegetation Data*: For comparing plots of vegetation, homogeneity is modulated by a density factor as follows (Hall, 1967):

$$H_{qm} = \sum_{j=1}^p \left(\sum_n a_j \right) \left(\sum_{j=1}^p \sum_n a_j \right)^{-1} \left(1 - \frac{s_{jqn}}{S_{qnm}} \right)$$

In the unsimplified form of the function, the homogeneity of the j th of the p attributes (taxa) is weighted by the average of the abundance values for that attribute for the n sites, divided by the average of such values for all the attributes. The abundances are all recorded on the same scale with zero representing absence.

Data for the test study were taken from a belt transect in South-West Cape vegetation at Happy Valley, Bains Kloof, near Wellington. The transect passed from river-bank vegetation (plot 16 in Fig. 5) through an intermediate zone (plot 15), across a sandy plain (plots 9–14), up a steepish sandy slope (plots 5–8) followed by a dry, rocky, more gradual slope (plots 1–4). The sixteen $1\text{ m} \times 4\text{ m}$ plots lie end to end along the transect. The abundances of 54 species on the transect were included in the study.

The need for first grouping the plots, which are in effect samples of vegetation types, into classes was illustrated in the first test attempted. The plots themselves were grouped by homogeneity analysis. All but the last links seemed reasonably appropriate. At the end however, the river-bank and river-bank margin plots (16, 15) took precedence over the four upper slope plots (1–4) in adding to a large cluster formed by the remainder. Plots 15 and 16 clearly carry the more different vegetation, however. The reason for this fallacious result is shown in a more exaggerated case in Fig. 1 of this paper.

A Group Average Linkage dendrogram was therefore prepared for setting the plots into classes, using H_{qm} values for pair links alone. This is shown in Fig. 5. Five classes are evident: plots 1–4, 5–8, 9–14, 15, 16. In a similar test of the data using the Czekanowski Coefficient (Czekanowski, 1913; Curtis, 1959) the linking sequence was broadly similar. Although the river-bank and river-bank margin plots were well separated from the remainder, other subdivisions were much less distinct, however.

The abundances of each species were then averaged in the plots of each class. A Homogeneity Analysis of this data gave the generally satisfactory dendrogram shown in Fig. 6. Further tests on data from the same area are in hand.

PROGRAMS AND COMPUTATION

Computer programs for Heterogeneity Analysis were written in MAC (Manchester AutoCode). The number of instructions needed ranged from about two hundred (two-state function, H_t) to 356 (quantitative and two-state function, H_{tq}).

After reading in the data, memory stores for the working matrix are set to a large number, half of which are subsequently replaced by a set of heterogeneity values for pair comparisons. Quantitative data are reduced, for each attribute, to a range with a maximum of 1 in the read-in part of the program, so that pair comparisons can be made using the difference of these new values for each

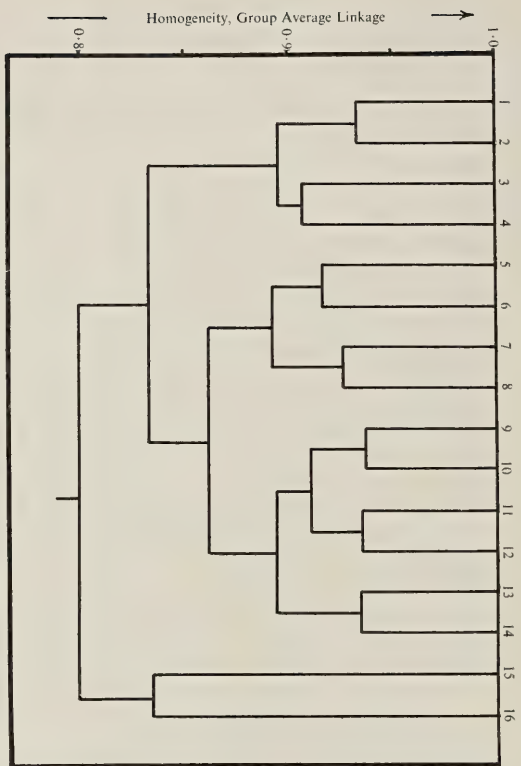


Fig. 5. Dendrogram showing the results of applying the abundance-sensitive Homogeneity Function $H_q^{(H)}$ with pair-comparisons and Group Average Linkage, to data from a belt transect in Cape vegetation.

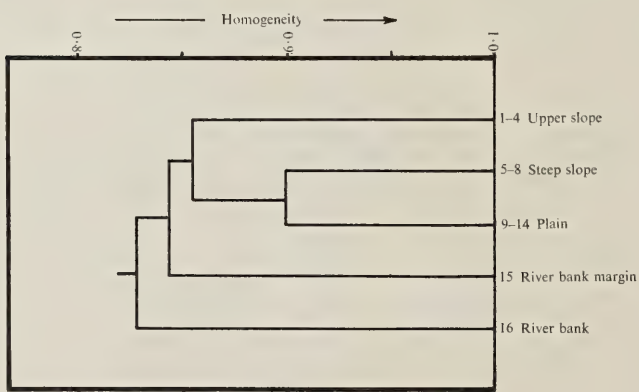


Fig. 6. Dendrogram showing the results of applying Homogeneity Analysis to the classes formed in Figure 5.

attribute, avoiding the standard deviation calculation for the large pair-link matrix. This results in a considerable saving in machine time. The denominator of the standard deviation ratio is also more simply calculated.

After preparing the first matrix of heterogeneity values, the group with the smallest value is found and recorded by the printer. After replacing each of the heterogeneity values relating to the two members linked, by a large number, values are calculated for the trial links of the new group with other members, using in the case of quantitative data the full standard deviation formula. These are again set in the working matrix so that the next minimum heterogeneity case may be determined.

Alternative pathways are detected by the program, and up to any four may be selected by manual control. Alternative pathways, and occasional small reversals, have been found so far only in the tests with two-state data.

Allowance for absence of some of the data about the members has been programmed so far only for the quantitative case, small-scale tests indicating satisfactory results.

Machine times, using the I.C.T. 1301 computer at the University of Cape Town, were within reasonable limits for larger groups. For the program for the two-state function, 40 members described by 93 attributes took 74 minutes 50 seconds; for six members and twenty attributes the time was 56 seconds. For quantitative data, these times would be approximately doubled.

CONCLUSIONS

Tests in this study show that, for taxa and vegetation with which the author has had experience, Heterogeneity Analysis provides generally satisfactory groupings. In view of the range of cases that can be accommodated by Heterogeneity Analysis, and its apparent lack of theoretical objections, it would seem to have a promising future. Further testing is needed, especially of the systems for interpreting dendrograms and the application of the method for ecology.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the helpful assistance in this work of the following persons at the University of Cape Town: Mr. G. B. Brundrit of the Department of Applied Mathematics and Mr. J. Field of the C.S.I.R. Oceanographic Research Unit, for help with principles and procedures as well as the use of their program for the Czekanowski Coefficient; Mr. Alan Zinober, student in the Faculty of Engineering, for writing the main parts of the computer programs; Mrs. E. P. Powrie and students in the 1966 Ecology Field Course for providing the belt transect data. The author also wishes to acknowledge the support of the Fourcade Bequest in meeting publication costs.

REFERENCES

- CURTIS, J. T. (1959). The vegetation of Wisconsin. Wisconsin.
CZEKANOWSKI, J. (1913). Zarys metod statystycznych. Warsaw.
HALL, A. V. (1965). Studies of the South African species of *Eulophia*. *Journ. S. Afr. Bot.* Suppl. Vol. 5.
HALL, A. V. (1967). Methods for demonstrating resemblance in taxonomy and ecology. *Nature* **214** : 830-831.
LANCE, G. N. AND WILLIAMS, W. T. (1966a). Computer programs for hierarchical polythetic classification ("similarity analyses"). *Computer Journ.* **9**: 60-64.
LANCE, G. N. AND WILLIAMS, W. T. (1966b). A generalized sorting strategy for computer classifications. *Nature* **212**: 218.
SOKAL, R. R. AND MICHENER, C. D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **38**: 1409-1438.
SOKAL, R. R. AND SNEATH, P. H. A. (1963). Principles of numerical taxonomy. W. H. Freeman & Co., San Francisco.